Capstone Project
The Battle of Neighborhoods

IBM Data Science Professional Certification

1. Introduction

- Moving to a new city is not easy because you want a similar or better infrastructure and quality of life.
- A comparison between the districts of the new area and the hometown would be very helpful to find out which are similar and can therefore be prioritized.
- Mr. Smith is a real estate agent and specializes in supporting clients who want to leave their hometown and move to another unfamiliar area.
- His new client is Mr. Miller who lost his job in New York and found a new job in Toronto. Now he would like to move there with his family. But the similarity of the cities is very important for him.
- This can also be applied to a wide range of applications. When a company, such as a restaurant decides to expand in a new country or to open a new branch. It would be advantageous for the company to find a similar neighborhood, because there are often interactive effects between the shops.

2. Description of Data
- Venues of FourSquare

1. Venues of FourSquare: a location-based recommendation service in the form of application software for event locations
   - total of 10 categories, which are divided into a total of 470 sub-categories

	Categorie	Amount of Subcategories
0	Arts & Entertainment	38
1	College & University	23
2	Event	12
3	Food	92
4	Nightlife Spot	7
5	Outdoors & Recreation	66
6	Professional & Other Places	44
7	Residence	5
8	Shop & Service	147
9	Travel & Transport	36

2. Description of Data
- Neighborhoods of New York

2. Neighborhoods and coordinates of New York: Mr. Miller's current home
   - total of 5 boroughs with a total of 306 neighborhoods
   - Mr. Miller's coordinates: latitude = 40.7127281, longitude = -74.0060152

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

2. Description of Data
- Neighborhoods of Toronto

3. Neighborhoods and coordinates of Toronto: Mr. Miller's future home
   - total of 10 boroughs, with a total of 103 neighborhoods
   - coordinates of his new job: latitude = 43.6534817, longitude = -79.3839347

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.8114	-79.1966
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.7857	-79.1587
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.7658	-79.1747
3	M1G	Scarborough	Woburn	43.7681	-79.2176
4	M1H	Scarborough	Cedarbrae	43.7694	-79.2389

3. Methodology
- Exploratory Data Analysis

- Determine the venues of neighborhoods within a radius of 1km.
- There are a total of 333 unique sub-categories in Toronto.

	Venue	Venue Category
Neighborhood		
Agincourt	43	43
Alderwood, Long Branch	27	27
Bathurst Manor, Wilson Heights, Downsview North	29	29
Bayview Village	7	7
Bedford Park, Lawrence Manor East	38	38
Berczy Park	100	100
Birch Cliff, Cliffside West	16	16
Brockton, Parkdale Village, Exhibition Place	100	100

- For example, “Malvern, Rouge” neighborhood has the following four unique subcategories: 'Zoo Exhibit', 'Fast Food Restaurant', 'Trail', 'Hobby Shop’
- There are a total of 5 entries of venues.

	Neighborhood	Venue	Venue Category
0	Malvern, Rouge	Canadiana exhibit	Zoo Exhibit
1	Malvern, Rouge	Wendy's	Fast Food Restaurant
2	Malvern, Rouge	Grizzly Bear Exhibit	Zoo Exhibit
3	Malvern, Rouge	Upper Rouge Trail	Trail
4	Malvern, Rouge	Lee Valley	Hobby Shop

3. Methodology
- Machine Learning

- To find similar neighborhoods, Toronto neighborhoods need to be grouped by venue.
- The machine learning algorithm clustering is used to form these clusters.
- This is an unsupervised learning algorithm.
- For clustering, the two most popular techniques are used to determine the feasibility of this problem.
- These are KMeans and Density-Based Spatial Clustering.

3. Methodology
- Machine Learning - DBSCAN

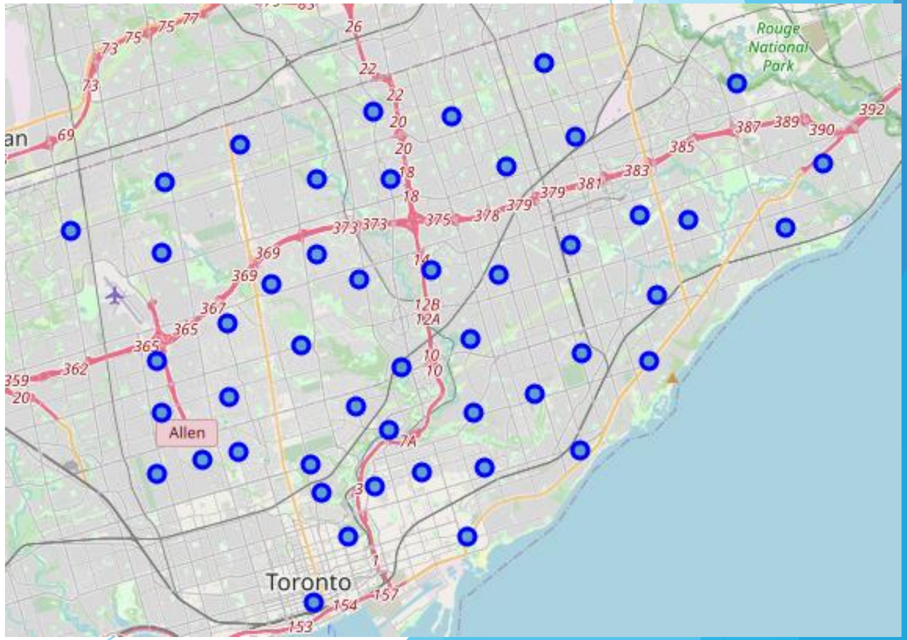
- DBSCAN is useful for studying spatial data.
- The algorithm creates clusters of arbitrary shape.
- Advantages:
  - The algorithm is relatively efficient for medium-sized and large data sets.
  - Can find clusters that are completely surrounded by another cluster.
  - It has an idea of noise and is robust to outliers.
  - It does not need any information about the number of clusters.
- Disadvantages:
  - It's a little slower than KMeans in terms of time and complexity.
  - It doesn't work well with clusters of different densities.

3. Methodology
- Machine Learning - KMeans

- KMeans is mainly used for segmenting customers.
- It groups the data in K non-overlapping subsets or clusters without a cluster-internal structure or labels.
- The intra-cluster distances are minimized and the inter-cluster distances are maximized.
- A local optimum is found.
- Advantages:
  - The algorithm is relatively efficient for medium and large data sets.
  - It creates sphere-like clusters as the clusters are shaped around the centroids.
- Disadvantages:
  - The number of clusters has to be specified beforehand.

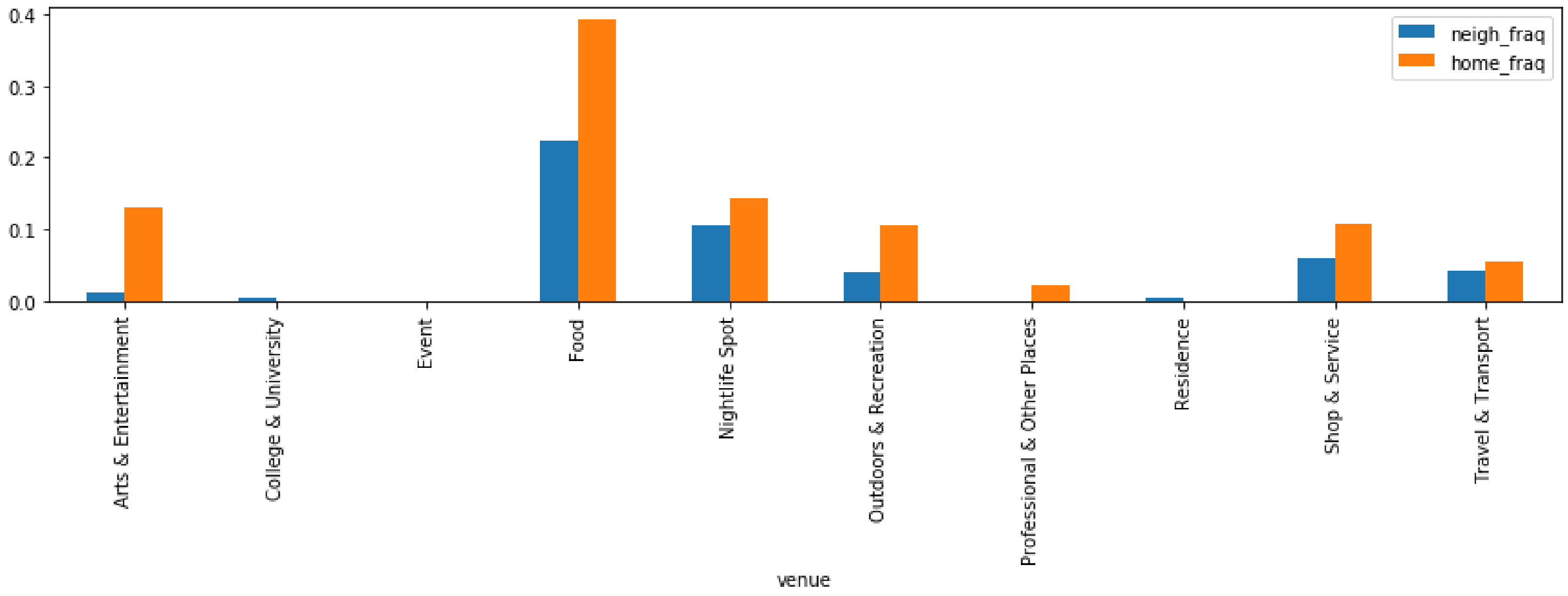
4. Results

- The DBSCAN could not group the data set because it only specified one cluster.
- KMeans has formed a total of 10 clusters.
- Mr. Miller's current neighborhoods of New York is most similar to the neighborhoods in cluster No. 7 in Toronto.
- 47 neighborhoods belong to this cluster.



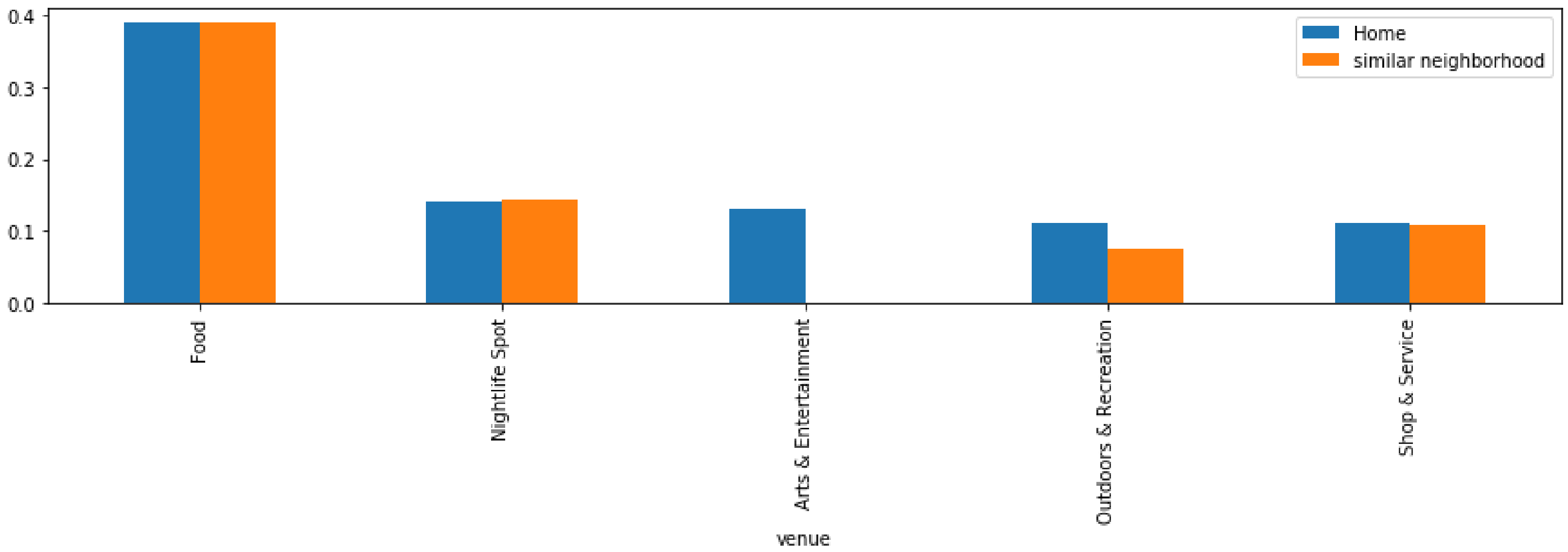
4. Results – Validation – all neighborhoods of most similar cluster

- Compare the frequency of the ten main categories.
- “Arts & Entertainment” and “Food” are much more often in Mr. Miller's current home than in the new area around Toronto.
- Otherwise, all other categories are relatively close to each other.

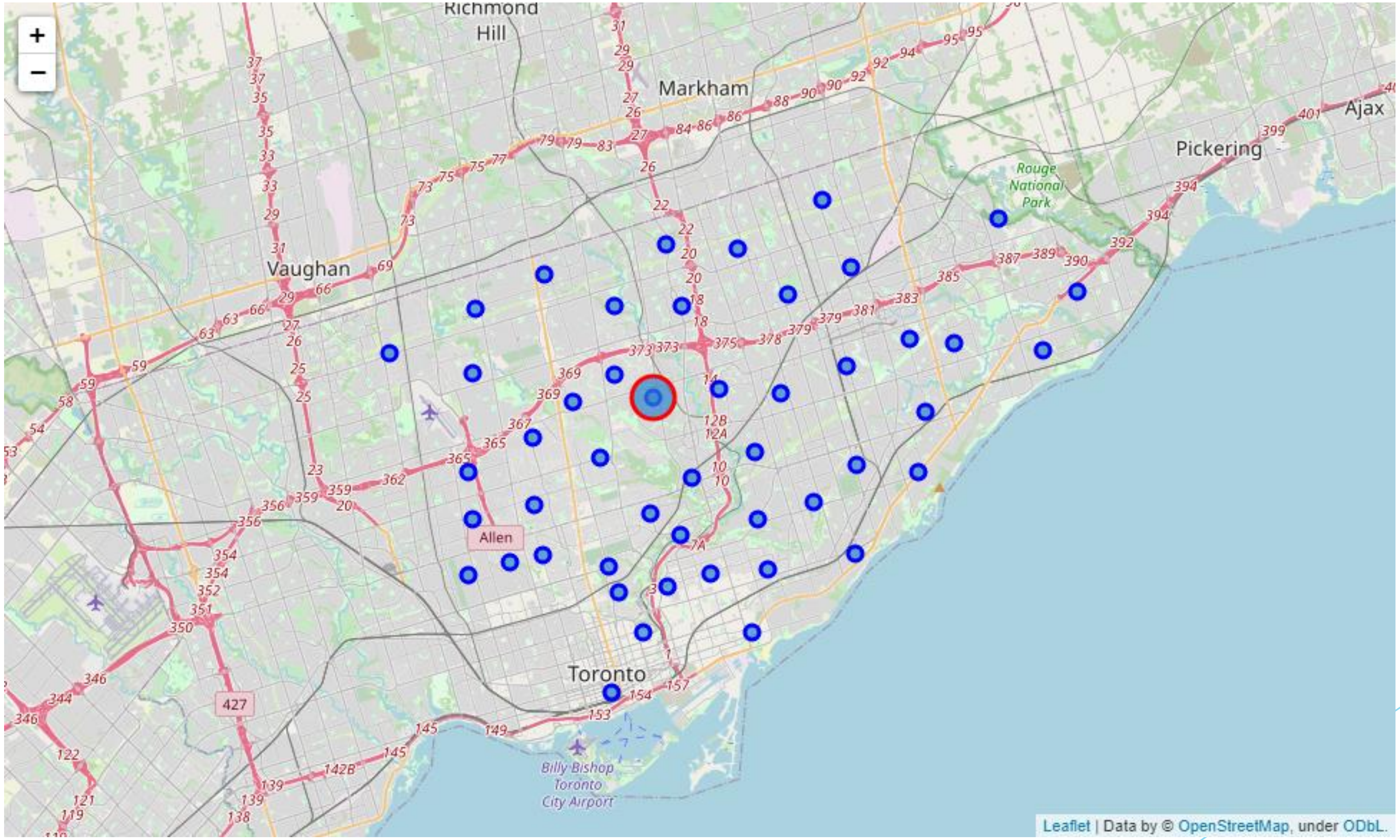


4. Results – Validation – most similar neighborhood

- Determine the frequency of the five most common venues.
- Neighborhood „Don Mills“ has the smallest mean absolute error.
- So, it is the closest neighborhood.



4. Results – most similar neighborhood



5. Discussion

- The results of an unsupervised learning model are sometimes difficult to assess.
- The result can also be validated with various methods, for example with validation in the sample or with descriptive statistics.
- DBSCAN is not robust enough for clusters with different densities over a high-dimensional data space.
- For this problem it is also advantageous to decide on a neighborhood, since this way the preferred wishes of the client can be better taken into account.

6. Conclusion

- The problem of finding a similar neighborhood is an important problem as it cannot only be applied to people who want to move.
- It can also be used for companies that want to open a new branch and are very satisfied in their current neighborhood so far.
- In this project, Toronto venues could be broken down into 10 main categories and 333 sub-categories.
- Two different cluster algorithms, DBSCAN and KMeans, were used to form clusters.
- KMeans clustering has established itself for this problem and has delivered better results.
- 10 clusters were formed, so that cluster No. 7 was most similar to the hometown. This cluster includes 47 neighborhoods.
- The most important event categories were verified and compared with those of the hometown. A neighborhood "Don Mills" was selected that was most similar.